Introduction and Problem Formulation

Businesses rely heavily on understanding what drives high-value transactions in order to improve pricing strategies, inventory planning, and sales channel optimization. In this project, I analyze a real-world restaurant sales dataset to identify factors associated with high sales transactions and to evaluate machine learning models for predicting sales performance.

The primary objective of this study is to predict whether a transaction results in high sales based on transaction level attributes such as price, quantity sold, purchase type, and city. Rather than predicting exact revenue amounts, I frame the problem as a binary classification task, where transactions are categorized as either high sales or low sales. This formulation aligns with real world decision making scenarios in which businesses seek to identify high-value transactions or customers.

By applying multiple classification models and comparing their performance, this project aims to determine which modeling approach provides the best balance between predictive accuracy and interpretability for this dataset.

Dataset Description

The dataset used in this project consists of 254 restaurant sales transactions and was obtained from a publicly available sales dataset. Each observation represents an individual transaction and includes information related to pricing, quantity sold, purchase characteristics, and location.

The dataset contains variables such as order identifier, transaction date, product name, price, quantity sold, purchase type, payment method, manager, and city. For the purposes of this analysis, I focus on variables that are most relevant to sales performance, including price, quantity, purchase type, and city.

During preprocessing, duplicate records were removed and rows with missing values were excluded to ensure data quality. I also engineered a new variable, total_sales, by multiplying price and quantity. Using this value, I created a binary target variable called high_sales, which indicates whether a transaction's total sales are greater than or equal to the median total sales value. This transformation resulted in a balanced classification target suitable for modeling.

This dataset is well suited for the analysis because it reflects realistic business scenarios and contains both numerical and categorical features that allow for meaningful exploratory analysis and model comparison.

Exploratory Data Analysis

I began the analysis by exploring the structure and distribution of the dataset to better understand patterns in sales behavior. Summary statistics were computed for numerical variables such as price, quantity sold, and total sales. The distribution of total sales was found to be right skewed, with most transactions concentrated around the median value and a small number of transactions exhibiting very high sales values. This indicates that while high value transactions are relatively rare, they contribute disproportionately to overall revenue.

To further examine sales patterns, I compared total sales across different purchase types using boxplots. In store purchases exhibited higher median total sales and greater variability compared to online and drive thru purchases. This suggests that customers purchasing in store are more likely to place larger orders, potentially due to impulse buying or bundled purchases.

I also analyzed the rate of high sales transactions across different cities. The results showed noticeable variation by location, with certain cities displaying a higher proportion of high sales transactions than others. This indicates that city level effects may influence purchasing behavior and that location is a meaningful feature for predicting high sales outcomes.

Overall, the exploratory analysis revealed that price, quantity, purchase type, and city all show relationships with sales performance. These findings informed the selection of predictor variables and supported framing the problem as a classification task focused on identifying high value transactions.

Model Selection and Methodology

The objective of this project is to predict whether a transaction results in high sales, which is a binary outcome. Therefore, I treat this problem as a classification task. Based on the exploratory analysis, I selected price, quantity sold, purchase type, and city as predictor variables, as these features showed meaningful relationships with sales performance. The target variable, high_sales, indicates whether a transaction's total sales exceed the median value.

I began modeling with logistic regression as a baseline classifier. Logistic regression is well suited for binary classification problems and offers strong interpretability, allowing for clear understanding of how each predictor influences the probability of high sales. This model serves as a reference point for evaluating the performance of more complex approaches.

To address potential overfitting and improve model stability, I applied regularized logistic regression models using ridge (L2) and lasso (L1) penalties. Ridge regularization shrinks coefficient magnitudes while retaining all predictors, whereas lasso regularization can shrink some coefficients to zero, effectively performing feature selection. These regularized models help control complexity and assess whether simpler or more constrained models improve generalization.

In addition to linear models, I included a decision tree classifier to capture potential non linear relationships and interactions between predictors. Decision trees do not rely on linear assumptions and can naturally model feature interactions, providing a useful comparison to logistic regression based methods.

The dataset was split into training and testing sets using an 80 percent training and 20 percent testing split. Stratification was applied to preserve the proportion of high and low sales transactions in both sets. Model performance was evaluated on the test set using classification accuracy, and results were compared across models to identify trade offs between predictive performance and interpretability.

Model Evaluation and Results

Model performance was evaluated on the held out test set using classification accuracy. Logistic regression was used as the baseline model, followed by ridge and lasso regularized logistic regression models and a decision tree classifier. Across all models, predictive performance was generally similar, indicating that the relationship between the predictors and the target variable is largely linear.

The baseline logistic regression model achieved strong performance, demonstrating that price, quantity, purchase type, and city provide sufficient information to distinguish between high and low sales transactions. Ridge and lasso regularized logistic regression models achieved comparable or slightly improved accuracy relative to the baseline, while offering improved control over model complexity. The lasso model additionally performed implicit feature selection by shrinking less informative coefficients to zero.

The decision tree classifier achieved very high accuracy on the test set. However, given the relatively small size of the dataset, this result may indicate overfitting rather than true generalization performance. This highlights the importance of comparing complex models with simpler and regularized alternatives.

Overall, regularized logistic regression models provided a strong balance between predictive performance and interpretability, making them well suited for this classification task.

Overall, regularized logistic regression models provided a strong balance between predictive performance and interpretability, making them well suited for this classification task.

Discussion and Real World Implications

The results of this analysis suggest that transaction level features such as price, quantity sold, purchase type, and city play important roles in determining whether a transaction results in high sales. The consistent performance of logistic regression and its regularized variants indicates that these relationships are largely linear and stable across the dataset.

From a business perspective, the findings highlight actionable insights. Higher quantities and in store purchases are associated with increased likelihood of high sales, suggesting that encouraging bundled purchases or enhancing the in store experience may lead to higher revenue. Additionally, variation in high sales rates across cities indicates that location specific strategies may be effective, such as targeted promotions or pricing adjustments in lower performing regions.

The use of regularized logistic regression demonstrates how businesses can build interpretable predictive models that balance accuracy with simplicity. Such models can be used to flag potentially high value transactions, support inventory planning, and guide sales strategy decisions without relying on overly complex modeling techniques.

Conclusion and Limitations

In this project, I developed a complete machine learning pipeline to predict whether a restaurant transaction results in high sales. Using a real world sales dataset, I performed exploratory data analysis, engineered relevant features, and evaluated multiple classification models. Logistic regression served as a strong and interpretable baseline, while ridge and lasso regularization improved model stability and helped manage complexity. A decision tree model was also explored to capture non linear relationships, though its performance suggested potential overfitting given the dataset size.

So, I can say that the results indicate that price, quantity sold, purchase type, and city are meaningful predictors of high sales transactions. Among the evaluated models, regularized logistic regression offers the best balance between predictive performance and interpretability, making it a practical choice for real world business applications.